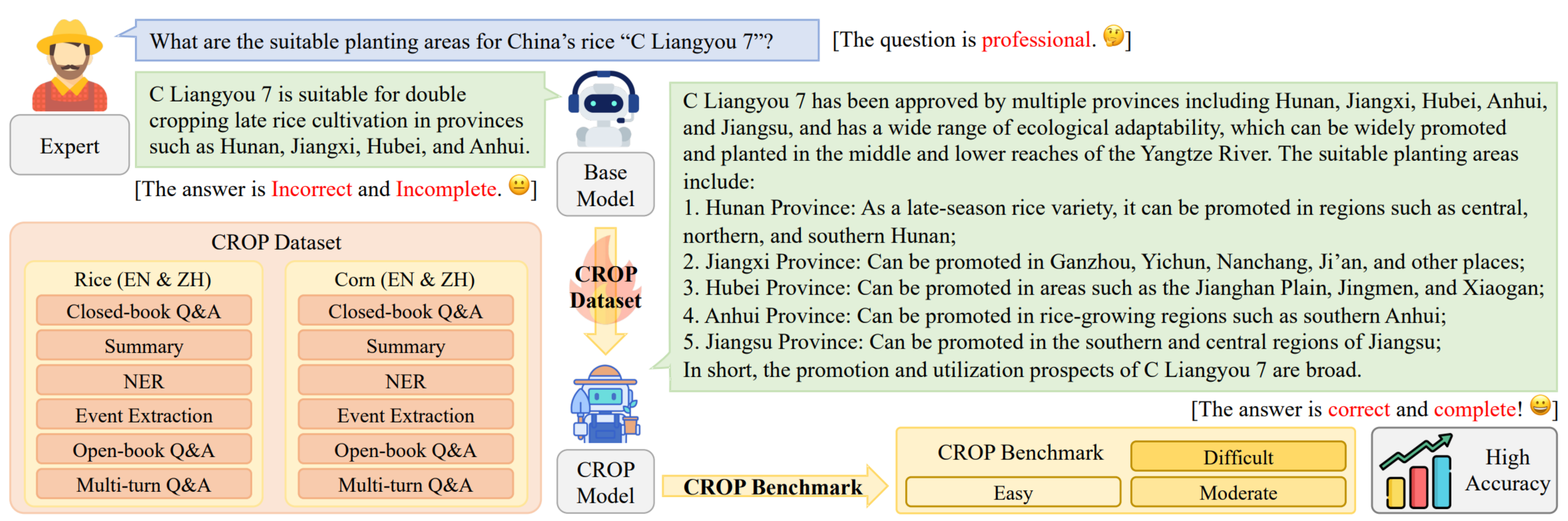
Empowering and Assessing the Utility of Large Language Models in Crop Science

Hang Zhang[1]*, Jiawei Sun[1]*, Renqi Chen[1]*, Wei Liu[1], Zhonghang Yuan[1], Xinzhe Zheng[1], Zhefan Wang[1], Zhiyuan Yang[4], Hang Yan[1], Hansen Zhong[1], Xiqing Wang[3], Wanli Ouyang[1], Fan Yang[2]†, Nanqing Dong[1]†
[1] Shanghai Artificial Intelligence Laboratory [2] Yazhouwan National Laboratory [3] China Agricultural University [4] Hangzhou Dianzi University

Github:

1. Motivation for the CROP

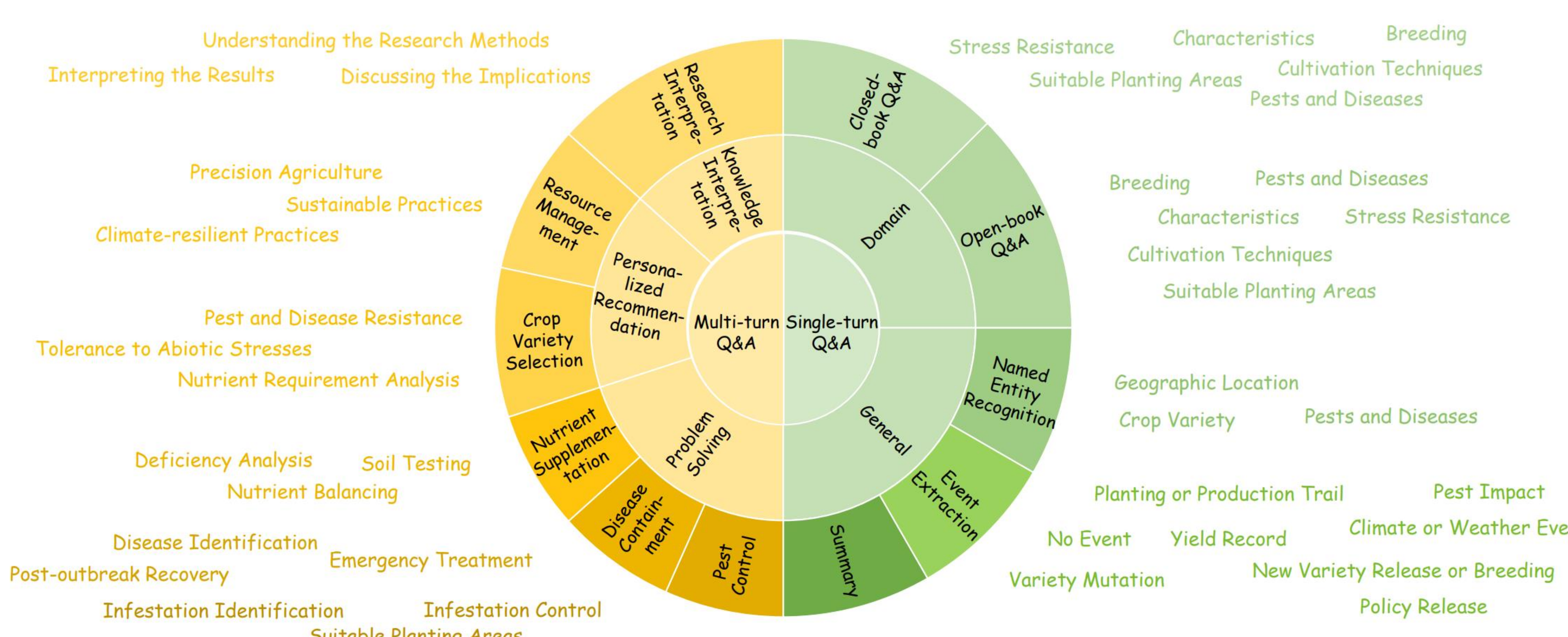
- Crop cultivation has historically been a significant challenge, with uncertainties in harvest yields.
- Recent progress in large language models (LLMs), offers promising opportunities. LLMs can generate professional knowledge and context in response to user inquiries, finding applications in various fields.
- However, LLMs currently face limitations in specific areas, such as pest management, and the existing datasets for agricultural evaluation are insufficient in quantity and locality.



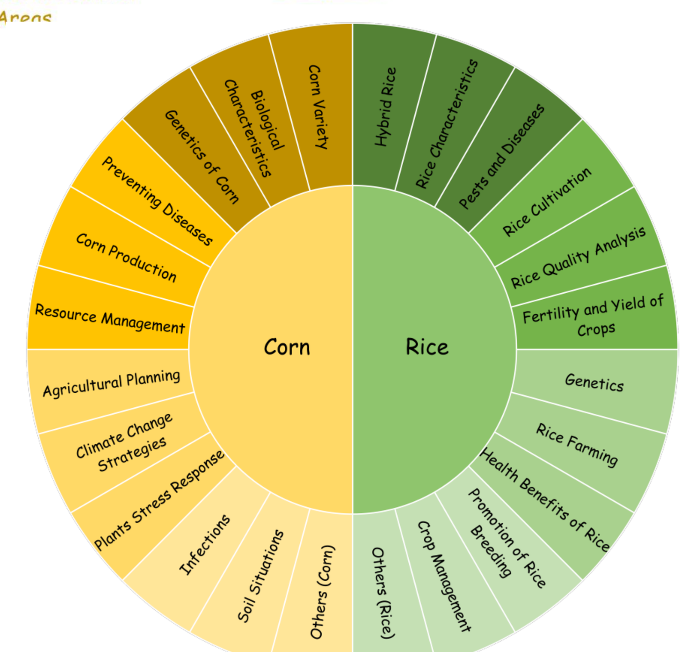
2. Overview of the CROP

To harness the full potential of LLMs for crop science, we propose a suite called CROP, which encompasses

- an extensive instruction-tuning dataset, designed to enhance the domain-specific proficiency of LLMs in crop science.
- a meticulously constructed benchmark, aimed at assessing the performance of LLMs across a variety of domain-related tasks.

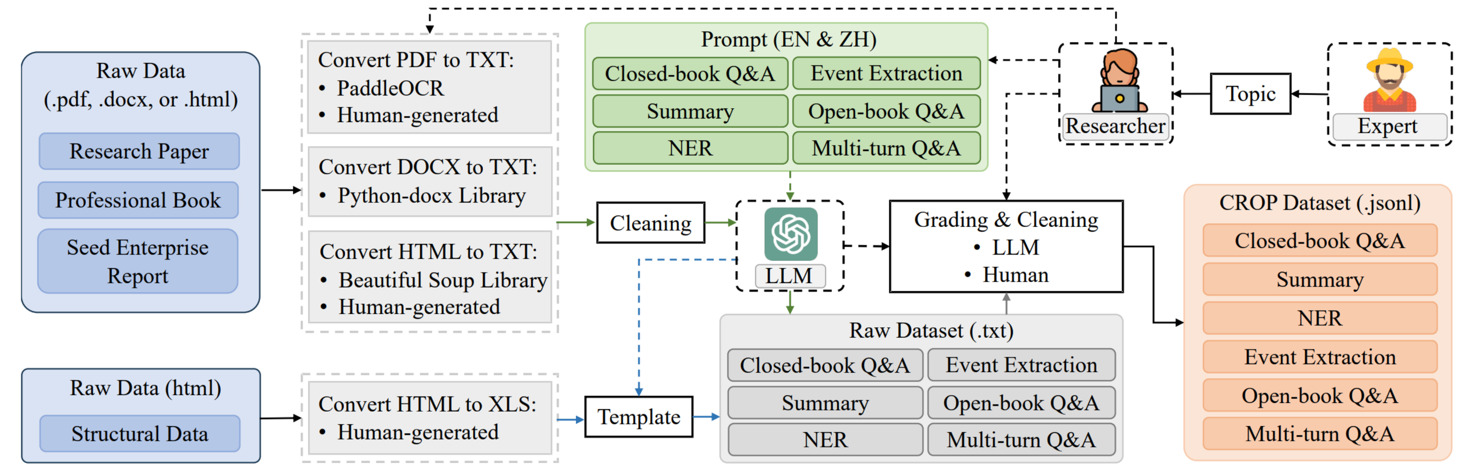


CROP dataset



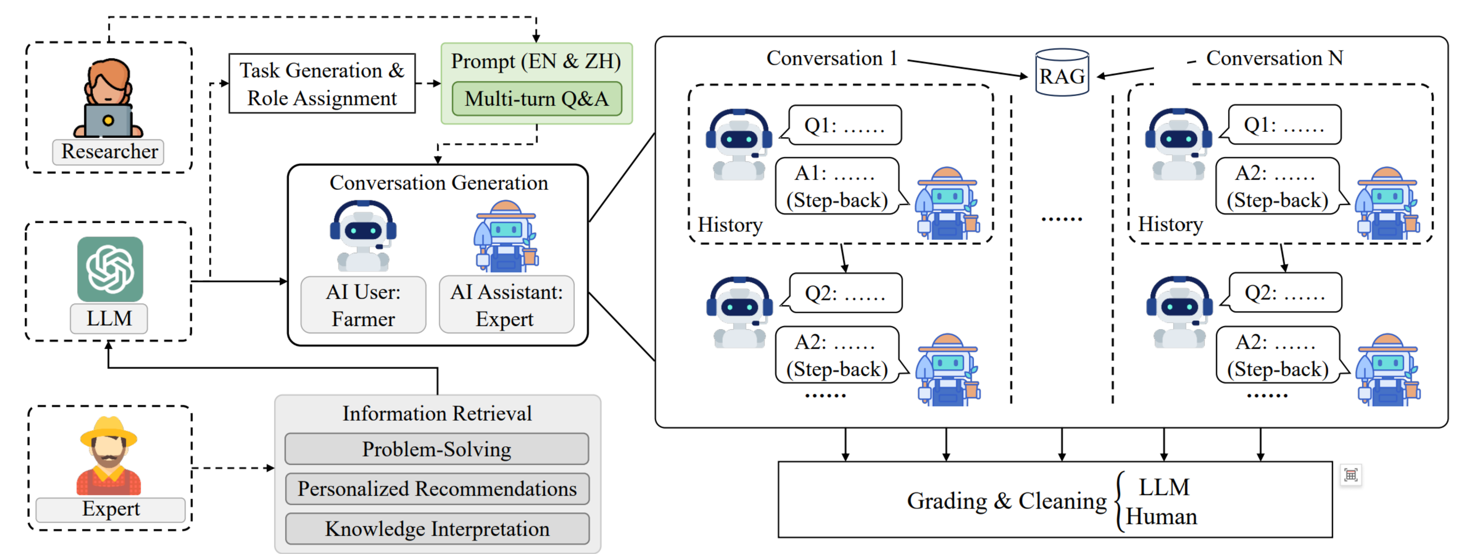
CROP benchmark

3. CROP Dataset Collection



Schematic overview of the dialogue collection procedure

- Raw data is first converted to TXT or XLS format.
- Prompt an LLM to generate Q&As from unstructured data or design templates that transform structured data into dialogue format.
- Filtering steps with both human and LLM involved.



Schematic overview of the multi-turn dialogue dataset collection procedure

- An LLM creates tasks under the guidance of domain experts and assigns roles to two agents.
- Using task-dependent prompts from researchers, the LLM generates dialogues with RAG.
- Filtering steps.

4. CROP Dataset Analysis

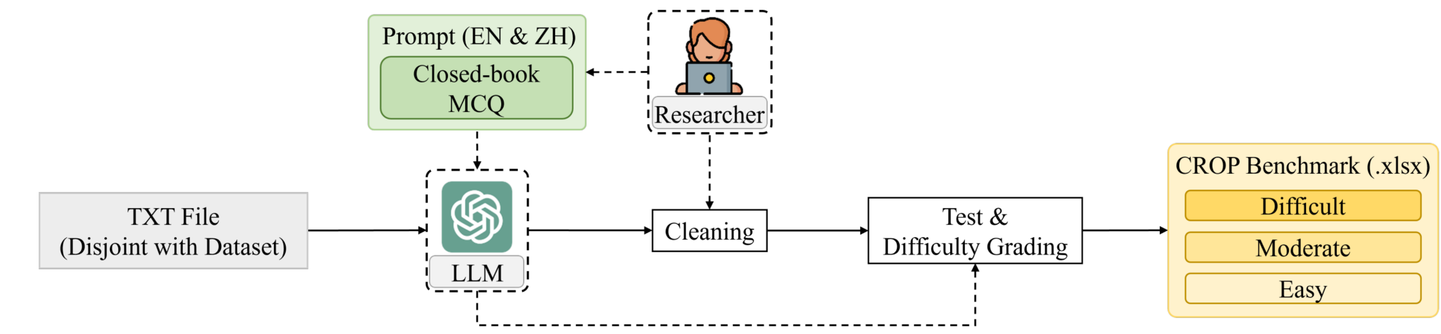
Composition of single-turn dialogues

Cereal	Type	Task	Abbr.	English Q&A	Chinese Q&A	Total
Rice	Domain	Closed-book Q&A	CQA	42,951	83,396	126,347
		Open-book Q&A	OQA	2,430	2,037	4,467
	General	Event Extraction	EE	1,891	1,030	9,742
		Named Entity Recognition	NER	2,003	1,604	
		Summary	Summary	1,586	1,628	
Corn	Domain	Closed-book Q&A	CQA	25,259	27,667	52,926
		Open-book Q&A	OQA	3,202	3,047	6,249
	General	Event Extraction	EE	2,245	1,322	10,307
		Named Entity Recognition	NER	2,008	1,316	
		Summary	Summary	1,559	1,857	
Others*		—				<1000
Overall		—		85,134	124,904	**210,038**

Composition of multi-turn dialogues

Cereal	Scenario	Task	English Q&A	Chinese Q&A	Total
Rice	Problem Solving	Pest Control	14+71	8+37	130
		Nutrient Supplementation	19+93	2+ 90+ 1	205
		Disease Containment	19+ 60	4+ 39	122
	Personalized Recommendation	Crop Variety Selection	12+ 53	9+ 9	83
		Resource Management	4+ 110+ 1	5+ 50	170
	Knowledge Interpretation	Research Interpretation	3+ 125+ 1	8+ 85	222
Corn	Problem Solving	Pest Control	20+ 84	7+ 77	188
		Nutrient Supplementation	24+ 56	8+ 30	118
		Disease Containment	21+ 64	2+ 19+ 1	107
	Personalized Recommendation	Crop Variety Selection	19+ 75	46+ 47	187
		Resource Management	8+ 94	1+ 69	172
	Knowledge Interpretation	Research Interpretation	5+ 94+1	6+ 61	167
Overall	—		1,150	721	**1,871**

5. CROP Benchmark Collection



- We prompt an LLM to generate MCQs from TXT files.
- After additional filtering steps with both human and LLM involved, we get the CROP benchmark, comprising three difficulty levels.

6. CROP Benchmark Analysis

- 5,045 questions in the benchmark have three difficulty levels:
- ➢ Easy (1613, 31.97%)
- ➢ Moderate (2754, 53.72%)
- ➢ Difficult (722, 14.31%)
- CROP benchmark consists of 5045 Chinese and English MCQs and covers 22 countries across six continents.

7. Experiments

1. The performance of selected LLMs on the CROP benchmark

Model	Access	Size	Overall ↑	Difficulty		
				Easy ↑	Moderate ↑	Difficult ↑
Commercial LLMs						
GPT-4[1]	API	N/A	0.856	1.000[2]	1.000[2]	0.000[2]
GPT-3.5[1]	API	N/A	0.328	1.000[2]	0.000[2]	0.061
Claude-3[1]	API	N/A	0.900	0.982	0.968	0.458
Qwen[1]	API	N/A	0.866	0.987	0.945	0.301
Open-source LLMs						
LLaMA3-Base	Weights	8B	0.348	0.443	0.341	0.161
+CQIA	Weights	8B	0.643 (+0.295)	0.791 (+0.348)	0.651 (+0.310)	0.281 (+0.120)
+CROP	Weights	8B	0.752 (+0.404)	0.866 (+0.432)	0.772 (+0.431	0.378 (+0.217)
+CQIA+CROP	Weights	8B	0.754 (+0.406)	0.918 (+0.475)	0.779 (+0.438)	0.295 (+0.134)
Qwen1.5-Base	Weights	7B	0.646	0.799	0.646	0.302
+CQIA	Weights	7B	0.688 (+0.042)	0.880 (+0.081)	0.689 (+0.043)	0.258 (-0.044)
+CROP	Weights	7B	0.676 (+0.030)	0.849 (+0.050)	0.688 (+0.042)	0.202 (-0.100)
+CQIA+CROP	Weights	7B	0.709 (+0.063)	0.910 (+0.111)	0.704 (+0.058)	0.227 (-0.075)
InternLM2-Base	Weights	7B	0.368	0.445	0.381	0.148
+CQIA	Weights	7B	0.723 (+0.355)	0.861 (+0.416)	0.750 (+0.369)	0.317 (+0.169)
+CROP	Weights	7B	0.748 (+0.380)	0.945 (+0.500)	0.761 (+0.380)	0.212 (+0.064)
+CQIA+CROP	Weights	7B	0.768 (+0.400)	0.939 (+0.494)	0.794 (+0.413)	0.285 (+0.137)

2. The performance of fine-tuned LLMs under different training epochs and languages.

Model	Epoch	Size	Overall ↑	Difficulty			Language		
				Easy ↑	Moderate ↑	Difficult ↑	Chinese ↑	English ↑	Variation ↓
LLaMA3-Base	N/A	8B	0.348	0.443	0.341	0.161	0.327	0.369	4.2%
+CQIA+CROP	1	8B	0.738	0.903	0.758	0.292	0.719	0.757	3.8%
+CQIA+CROP	2	8B	0.742	0.902	0.772	0.271	0.729	0.755	**2.6%**
+CQIA+CROP	4	8B	**0.754**	**0.918**	**0.779**	**0.295**	**0.738**	**0.770**	3.2%
Qwen1.5-Base	N/A	7B	0.646	0.799	0.646	**0.302**	0.667	0.624	4.3%
+CQIA+CROP	1	7B	0.702	**0.910**	**0.717**	0.183	**0.725**	0.680	4.5%
+CQIA+CROP	2	7B	0.670	0.875	0.677	0.181	0.690	0.649	4.1%
+CQIA+CROP	4	7B	**0.709**	**0.910**	0.704	0.227	0.717	**0.686**	**3.1%**
InternLM2-Base	N/A	7B	0.368	0.445	0.381	0.148	0.409	0.327	8.2%
+CQIA+CROP	1	7B	0.764	**0.942**	0.787	0.276	0.770	0.757	3.3%
+CQIA+CROP	2	7B	**0.809**	0.909	**0.855**	**0.414**	**0.811**	**0.807**	**0.4%**
+CQIA+CROP	4	7B	0.768	0.939	0.794	0.285	0.770	0.766	**0.4%**